

**AN AUTOMATIC CORRELATION METHOD FOR GENERATING SUMMARIES  
FOR TEXT DOCUMENTS**

**BACKGROUND OF THE INVENTION**

*Field of the Invention*

This invention relates to an automatic text processing method, and in particular, to a method for generating summaries for text documents.

*Background Description*

In information query, for the user's convenience, it is normally required that the summaries are generated for the users by means of automatic text processing functions of the computers. The current practical methods for automatically generating summaries for text documents are following four kinds:

- List the first paragraph of the document or the beginning paragraphs of the document as the summary (Infoseek, Yahoo, etc.): This method is simple, but does not suit for common document style;
- List the sentences in the query commands (Lotus website, Beijing Daily Online, etc.): The listed sentences relate directly to the query, but cannot represent the overall style of the document;
- Use implicit template: This method matches some patterns in the document, and then fills the matched contents into the pre-formed template. This method can generate very fluent summaries, but can only be suitable for a fixed document style and a specific domain, and is very difficult to be used commonly;
- Count the occurrence frequency of words or characters: This is a statistics-based method, which generally can be divide into four steps: (1) analyze the document discourse, and segment the document into paragraphs and sentences; (2) segment the sentences into words; (3) evaluate the scores of the importance of the words and the

sentences; (4) output the sentences with higher evaluated scores as the document's summary.

Although the above statistics-based method for automatically generating summaries for text documents has considered the occurrence frequency of words and characters in documents and therefore evaluated the importance of the words and the sentences, the summaries can not well correspond to the user's requirements because there is no interaction with the user. Therefore, the invention proposes a method for automatically generating summaries for text documents, which, when receiving the user's text documents, queries the fields, topics, and terms that the user is interested in. The method extracts the important sentences, and then in reasonable order, outputs them as the document's summary. The method can not only generate summaries for respective documents, but also generate a comprehensive prompt for the important ideas of the documents.

### SUMMARY OF THE INVENTION

The method for automatically generating summaries for text documents according to the invention includes steps of:

For a set of documents, generating a set of sentences by document discourse analysis, and obtaining a set of words by morphologic processing;

Initializing a score for each word in the set of words, and each sentence in the set of sentences;

Computing the score for each word in the set of sentences according to the scores of sentences containing it and the correlation degree between the word and the user information;

Computing the score for each sentence in the set of sentences according to the scores of words composing it and the position of the sentence in a section and a paragraph;

If the sum of scores of the words and the sum of scores the sentences change apparently, go back to the step of computing the word scores, otherwise continuing.

Outputting the top-ranked sentences as the summary of the set of documents, the top-ranked words as keywords list of the set of documents.

The above method introduces the following functions into the traditional statistics-based methods:

- It has a new sentence ranking strategy called “automatic correlation method”, which is much responsive to user’s requirements;
- It supports user summarization profile, which allows a user to customize the fields, topics, and terms that the user is interested in.
- It applies to general-purpose, and is also suited for summarizing the certain query documents.

The method considers the following factors when computing the scores of words in a word set: the correlation degree between the word and the user’s summarization profile language; the similarity degree between the word and the query term or topic provided by the user; sum scores of the sentences to which the word belongs; the similarity degree between the word and the word terms in the document title; the ratio of the occurrence times of the word in the document to its occurrence times in the document set; and the ratio of number of the documents in which the word occurs, to the total number of the documents contained in the document set.

### **BRIEF DESCRIPTION OF THE DRAWINGS**

The advantages and features of the invention will be more apparent by the description of preferred embodiments of the invention in conjunction with the accompanying figures, in which:

Fig. 1 is a flow chart of the method for automatically generating summaries for text documents according to a particular embodiment of the invention; and

Fig. 2 is a flow chart used in the sentence ranking part in Fig. 1.

### **DESCRIPTION OF THE PREFERRED EMBODIMENTS**

As shown in Fig. 1, a method for automatically generating summaries for text documents according to a particular embodiment of the invention includes steps of:

Step 1. Document Discourse Analysis

This step identifies titles, sections, lists, paragraph boundaries and sentence boundaries of the documents.

### Step 2. Morphologic Process

This step performs morphologic process for each sentence according to the language of documents. For Chinese language, the morphologic process includes the steps of: (1) segmenting sentences into words according to the system dictionary and the user-defined dictionaries; (2) identifying proper names (currently including person names, place names and person titles), domain terms, numbers, measure words, and date expressions; (3) adding part-of-speech tags to each word; (4) resolving the pronominal anaphora of the personal pronouns; (5) identifying the word relationship (such as same object names, synonyms, concept relationships, etc.) and building a relationship network among all of the words, according to the system thesaurus and user thesaurus. For English language, this step includes the steps of: (1) normalizing terms to a standard term; (2) identifying proper names; (3) splitting compound terms; and (4) filtering stop-word.

### Step 3. Sentence Ranking

This step is to determine the importance of each sentence by an automatic correlation algorithm. This step is described in more details below.

### Step 4. Summary Output:

- If the user requires one summary for one document, then this step selects the top-ranked sentences to output according to their appearing order in this document;
- If the user requires a single comprehensive summary for the document set, then this step outputs the sentences according to their computed scores from high to low and marks the sentences to show from which document they come (for example, by adding hyperlinks to the sentences), so that the user can easily look up the respective document.

In both cases the pronouns will be replaced by their entities.

After the document discourse analysis and the morphologic process are performed for each document in the document set D, each sentence in the document set is ranked to decide its important degree according to the sentence set S and the word set W of each document. The sentences are ranked to compute their scores from each other by using an

autonomous correlation method, i.e. by using the sentence set S and the word set W. This is realized by the steps below (see Fig. 2):

Step 1. Introducing a data group SCORE to record the computed scores of the sentences and the words, and initializing a score SCORE[s] of each sentence and a score SCORE[w] of each word into 0;

Step 2. Computing a score SCORE[w] of each word according to the weighted-average of the following values:

The correlation degree between W and the user's summarization profile language;

The similarity degree between W and the query terms or topic provided by the user;

Sum scores of the sentences to which W belongs;

The similarity degree between W and the word terms in each document title;

The ratio of the occurrence times of W in the document to its occurrence times in the document set; and

The ratio of number of the documents in which W occurs, to the total number of the documents contained in the document set D;

This can be written by the following formula, i.e.

$$\begin{aligned} \text{SCORE}[w] = & \lambda_1 * \text{salience}(w, \text{user summarization profile}) \\ & + \lambda_2 * \text{salience}(w, \text{user's query or topic}) \\ & + \lambda_3 * \sum(\text{SCORE}[s], s \in \omega) \\ & + \lambda_4 * \text{salience}(w, \text{title words}) \\ & + \lambda_5 * \text{FREQUENCY}(w/d) / \text{FREQUENCY}(w/D) \\ & + \lambda_6 * \text{NUMBER}(d, d \ni w) / \text{NUMBER}(D) \end{aligned}$$

Formula 1

Step 3. Computing the SCORE[s] of the sentence according to the weighted-average of the three values below:

- Sum scores of all the words constituting the sentences;
- The position of the sentence in the paragraph and section: the first sentences in the paragraph and section get higher scores than the sentences in other positions;
- The similarity among the sentences: if in many documents there are sentences whose contents are similar, these sentences are weighted more;

This can be written as the following formula

$$\text{SCORE}[s] = \lambda_7 * \sum(\text{SCORE}[w], s \in w) + \lambda_8 * \text{position}(s, d) + \lambda_9 * \text{similarity}(s, S)$$

**Formula 2**

Step 4. If the sum of the all scores changes significantly, cycle Step 2; otherwise the process ends.

It is seen according to the description of the invention in conjunction with the particular embodiments that the summary method of the invention is also a statistics-based method, which performs the discourse analysis and the morphologic process, and that the new functions of the method are:

- Allowing the user to customize “the user summary profile” in which the user can list the fields and topics he or she interested in, and also listed he or she is sensitive to which kind of word (such as person names, person titles, place names, numbers, amounts of money, dates, terms defined by the user own, etc.);
- The generated summaries being capable of being driven by subjects or the user’s query;
- A completely new sentence ranking strategy, herein called “automatic correlation method”: the first step, initializing the ranking score for the words and sentences; the second step, computing scores for every words according to the user summarization profile, topics or query terms provided by the user, and frequencies of the words; the third step, computing the ranking scores according to the words contained in the sentences and the document discourse in the document set; the fourth step, feedbacking the sentence scores to the words and repeating the second step and the third step, until the sentence scores have been stabilized.

This method fully utilizes the discourse information of every document, the clue words in the document, the title words, the language processing results, and topics or query terms provided by the user, to make the generated summaries accord with the user’s requirements more completely.

The flow charts described here are only exemplary, and many modifications can be made to those chart examples or the steps (or operations) described therein without departing from the spirit of the invention. For example, those steps can be executed in different order, or increased, reduced or improved. Therefore, those changes are considered as a part of the invention which points out the claims.

Although the preferred embodiments have been described here, it is apparent to those skilled in the art that various of modifications, complements, replacements and similar changes can be made without departing from the spirit of the invention, therefor those alternations are considered to be in the inventive scope defined by the appended claims.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
3310  
3311  
3312  
3313  
3314  
3315  
3316  
3317  
3318  
3319  
33100  
33101  
33102  
33103  
33104  
33105  
33106  
33107  
33108  
33109  
33110  
33111  
33112  
33113  
33114  
33115  
33116  
33117  
33118  
33119  
331100  
331101  
331102  
331103  
331104  
331105  
331106  
331107  
331108  
331109  
331110  
331111  
331112  
331113  
331114  
331115  
331116  
331117  
331118  
331119  
3311100  
3311101  
3311102  
3311103  
3311104  
3311105  
3311106  
3311107  
3311108  
3311109  
3311110  
3311111  
3311112  
3311113  
3311114  
3311115  
3311116  
3311117  
3311118  
3311119  
33111100  
33111101  
33111102  
33111103  
33111104  
33111105  
33111106  
33111107  
33111108  
33111109  
33111110  
33111111  
33111112  
33111113  
33111114  
33111115  
33111116  
33111117  
33111118  
33111119  
331111100  
331111101  
331111102  
331111103  
331111104  
331111105  
331111106  
331111107  
331111108  
331111109  
331111110  
331111111  
331111112  
331111113  
331111114  
331111115  
331111116  
331111117  
331111118  
331111119  
3311111100  
3311111101  
3311111102  
3311111103  
3311111104  
3311111105  
3311111106  
3311111107  
3311111108  
3311111109  
3311111110  
3311111111  
3311111112  
3311111113  
3311111114  
3311111115  
3311111116  
3311111117  
3311111118  
3311111119  
33111111100  
33111111101  
33111111102  
33111111103  
33111111104  
33111111105  
33111111106  
33111111107  
33111111108  
33111111109  
33111111110  
33111111111  
33111111112  
33111111113  
33111111114  
33111111115  
33111111116  
33111111117  
33111111118  
33111111119  
331111111100  
331111111101  
331111111102  
331111111103  
331111111104  
331111111105  
331111111106  
331111111107  
331111111108  
331111111109  
331111111110  
331111111111  
331111111112  
331111111113  
331111111114  
331111111115  
331111111116  
331111111117  
331111111118  
331111111119  
3311111111100  
3311111111101  
3311111111102  
3311111111103  
3311111111104  
3311111111105  
3311111111106  
3311111111107  
3311111111108  
3311111111109  
3311111111110  
3311111111111  
3311111111112  
3311111111113  
3311111111114  
3311111111115  
3311111111116  
3311111111117  
3311111111118  
3311111111119  
33111111111100  
33111111111101  
33111111111102  
33111111111103  
33111111111104  
33111111111105  
33111111111106  
33111111111107  
33111111111108  
33111111111109  
33111111111110  
33111111111111  
33111111111112  
33111111111113  
33111111111114  
33111111111115  
33111111111116  
33111111111117  
33111111111118  
33111111111119  
331111111111100  
331111111111101  
331111111111102  
331111111111103  
331111111111104  
331111111111105  
331111111111106  
331111111111107  
331111111111108  
331111111111109  
331111111111110  
331111111111111  
331111111111112  
331111111111113  
331111111111114  
331111111111115  
331111111111116  
331111111111117  
331111111111118  
331111111111119  
3311111111111100  
3311111111111101  
3311111111111102  
3311111111111103  
3311111111111104  
3311111111111105  
3311111111111106  
3311111111111107  
3311111111111108  
3311111111111109  
3311111111111110  
3311111111111111  
3311111111111112  
3311111111111113  
3311111111111114  
3311111111111115  
3311111111111116  
3311111111111117  
3311111111111118  
3311111111111119  
33111111111111100  
33111111111111101  
33111111111111102  
33111111111111103  
33111111111111104  
33111111111111105  
33111111111111106  
33111111111111107  
33111111111111108  
33111111111111109  
33111111111111110  
33111111111111111  
33111111111111112  
33111111111111113  
33111111111111114  
33111111111111115  
33111111111111116  
33111111111111117  
33111111111111118  
33111111111111119  
331111111111111100  
331111111111111101  
331111111111111102  
331111111111111103  
331111111111111104  
331111111111111105  
331111111111111106  
331111111111111107  
331111111111111108  
331111111111111109  
331111111111111110  
331111111111111111  
331111111111111112  
331111111111111113  
331111111111111114  
331111111111111115  
331111111111111116  
331111111111111117  
331111111111111118  
331111111111111119  
3311111111111111100  
3311111111111111101  
3311111111111111102  
3311111111111111103  
3311111111111111104  
3311111111111111105  
3311111111111111106  
3311111111111111107  
3311111111111111108  
3311111111111111109  
3311111111111111110  
3311111111111111111  
3311111111111111112  
3311111111111111113  
3311111111111111114  
3311111111111111115  
3311111111111111116  
3311111111111111117  
3311111111111111118  
3311111111111111119  
33111111111111111100  
33111111111111111101  
33111111111111111102  
33111111111111111103  
33111111111111111104  
33111111111111111105  
33111111111111111106  
33111111111111111107  
33111111111111111108  
33111111111111111109  
33111111111111111110  
33111111111111111111  
33111111111111111112  
33111111111111111113  
33111111111111111114  
33111111111111111115  
33111111111111111116  
33111111111111111117  
33111111111111111118  
33111111111111111119  
331111111111111111100  
331111111111111111101  
331111111111111111102  
331111111111111111103  
331111111111111111104  
331111111111111111105  
331111111111111111106  
331111111111111111107  
331111111111111111108  
331111111111111111109  
331111111111111111110  
331111111111111111111  
331111111111111111112  
331111111111111111113  
331111111111111111114  
331111111111111111115  
331111111111111111116  
331111111111111111117  
331111111111111111118  
331111111111111111119  
3311111111111111111100  
3311111111111111111101  
3311111111111111111102  
3311111111111111111103  
3311111111111111111104  
3311111111111111111105  
3311111111111111111106  
3311111111111111111107  
3311111111111111111108  
3311111111111111111109  
3311111111111111111110  
3311111111111111111111  
3311111111111111111112  
3311111111111111111113  
3311111111111111111114  
3311111111111111111115  
3311111111111111111116  
3311111111111111111117  
3311111111111111111118  
3311111111111111111119  
33111111111111111111100  
33111111111111111111101  
33111111111111111111102  
33111111111111111111103  
33111111111111111111104  
33111111111111111111105  
33111111111111111111106  
33111111111111111111107  
33111111111111111111108  
33111111111111111111109  
33111111111111111111110  
33111111111111111111111  
33111111111111111111112  
33111111111111111111113  
33111111111111111111114  
33111111111111111111115  
33111111111111111111116  
33111111111111111111117  
33111111111111111111118  
33111111111111111111119  
331111111111111111111100  
331111111111111111111101  
331111111111111111111102  
331111111111111111111103  
331111111111111111111104  
331111111111111111111105  
331111111111111111111106  
331111111111111111111107  
331111111111111111111108  
331111111111111111111109  
331111111111111111111110  
331111111111111111111111  
331111111111111111111112  
331111111111111111111113  
331111111111111111111114  
331111111111111111111115  
331111111111111111111116  
331111111111111111111117  
331111111111111111111118  
331111111111111111111119  
3311111111111111111111100  
3311111111111111111111101  
3311111111111111111111102  
3311111111111111111111103  
3311111111111111111111104  
3311111111111111111111105  
3311111111111111111111106  
3311111111111111111111107  
3311111111111111111111108  
3311111111111111111111109  
3311111111111111111111110  
3311111111111111111111111  
3311111111111111111111112  
3311111111111111111111113  
3311111111111111111111114  
3311111111111111111111115  
3311111111111111111111116  
3311111111111111111111117  
3311111111111111111111118  
3311111111111111111111119  
33111111111111111111111100  
33111111111111111111111101  
33111111111111111111111102  
33111111111111111111111103  
33111111111111111111111104  
33111111111111111111111105  
33111111111111111111111106  
33111111111111111111111107  
33111111111111111111111108  
33111111111111111111111109  
33111111111111111111111110  
33111111111111111111111111  
33111111111111111111111112  
33111111111111111111111113  
33111111111111111111111114  
33111111111111111111111115  
33111111111111111111111116  
33111111111111111111111117  
33111111111111111111111118  
33111111111111111111111119  
331111111111111111111111100  
331111111111111111111111101  
331111111111111111111111102  
331111111111111111111111103  
331111111111111111111111104  
331111111111111111111111105  
331111111111111111111111106  
331111111111111111111111107  
331111111111111111111111108  
331111111111111111111111109  
331111111111111111111111110  
331111111111111111111111111  
331111111111111111111111112  
331111111111111111111111113  
331111111111111111111111114  
331111111111111111111111115  
331111111111111111111111116  
331111111111111111111111117  
331111111111111111111111118  
331111111111111111111111119  
3311111111111111111111111100  
3311111111111111111111111101  
3311111111111111111111111102  
3311111111111111111111111103  
3311111111111111111111111104  
3311111111111111111111111105  
3311111111111111111111111106  
3311111111111111111111111107  
3311111111111111111111111108  
3311111111111111111111111109  
3311111111111111111111111110  
3311111111111111111111111111  
3311111111111111111111111112  
3311111111111111111111111113  
3311111111111111111111111114  
3311111111111111111111111115  
3311111111111111111111111116  
3311111111111111111111111117  
3311111111111111111111111118  
3311111111111111111111111119  
33111111111111111111111111100  
33111111111111111111111111101  
33111111111111111111111111102  
33111111111111111111111111103  
33111111111111111111111111104  
33111111111111111111111111105  
33111111111111111111111111106  
33111111111111111111111111107  
33111111111111111111111111108  
33111111111111111111111111109  
33111111111111111111111111110  
33111111111111111111111111111  
33111111111111111111111111112  
33111111111111111111111111113  
33111111111111111111111111114  
33111111111111111111111111115  
33111111111111111111111111116  
33111111111111111111111111117  
33111111111111111111111111118  
33111111111111111111111111119  
331111111111111111111111111100  
331111111111111111111111111101  
3311111111111111